DOCUMENT RESUME

ED 340 746 TM 017 826

AUTHOR Chan, Jason C.

TITLE Estimating the Latent Trait from Likert-Type Data: A Comparison of Factor Analysis, Item Response Theory, and Multidimensional Scaling.

PUB DATE Aug 91

NOTE 35p.; Paper presented at the Annual Meeting of the American Psychological Association (99th, San Francisco, CA, August 16-20, 1991).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Comparative Analysis; Computer Simulation; Correlation; Equations (Mathematics); *Estimation (Mathematics); *Factor Analysis; *Item Response Theory; *Likert Scales; *Mathematical Models; *Multidimensional Scaling; Robustness (Statistics); Sample Size

IDENTIFIERS Graded Response Model; Pearson Product Moment Correlation; Skew Curves

ABSTRACT

The following seven statistical procedures are compared in terms of the ability to recover a unidimensional latent trait from Likert-type data: (1) factor analysis based on Pearson correlations (FA-PR); (2) factor analysis based on polychoric correlations (FA-PL); (3) the graded response model in item response theory (IRT-GRM); (4) internal unfolding (IMDU); (5) external unfolding (EMDU); (6) weighted unfolding (WMDU); and (7) summing up successive integers assigned to response categories (SSI). In this simulation study, sample size varied (n=30, n=100, and n=1,000 subjects), and test length (12 versus 24 items) and skewness of item response distributions were manipulated. Generally speaking, IRT-GRM performed best and was most robust against skewness. FA-PR and FA-PL performed equally well across almost all conditions, but were competitive with IRT-GRM only when item responses were normally distributed. SSI practice performed slightly worse than did the two FA procedures when item responses were normally distributed, but performed better when item responses were highly skewed. WMDU performed as well as did SSI only when item responses were normally distributed or moderately skewed and sample size was large for multidimensional scaling models (e.g., n=100). IMDU and EMDU performed worse than did WMDU and appeared to be inappropriate for Likert-type data. Eight tables present the comparative data. A 48-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED340746

Paper presented at the annual meeting of the American Psychological Association, San Francisco, 1991.

ESTIMATING THE LATENT TRAIT FROM LIKERT-TYPE DATA: A COMPARISON OF FACTOR ANALYSIS, ITEM RESPONSE THEORY, AND MULTIDIMENSIONAL SCALING

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
[x] This document has been reproduced as received from the person or organization originating it.
[ ] Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JASON CHINYU CHAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Jason C. Chan

The University of Texas at Austin

Correspondence:
Jason C. Chan
EDB 504
Department of Educational Psychology
The University of Texas at Austin
Austin, TX 78712

* The author would like to thank Dr. William R. Koch for his constructive criticisms and helpful suggestions. Comments from Dr. Barbara G. Dodd, Dr. William L. Hays, Dr. H. Paul Kelley, and Dr. John C. Loehlin were also acknowledged.

TM017826

BEST COPY AVAILABLE

Abstract

Seven statistical procedures were compared with one another in terms of the ability to recover a unidimensional latent trait from Likert-type data. They are factor analysis based on either Pearson correlations (FA-PR) or polychoric correlations (FA-PL), the graded response model in item response theory (IRT-GRM), the internal unfolding (IMDU), the external unfolding (EMDU), the weighted unfolding (WMDU), and the common procedure of summing up successive integers assigned to response categories (SSI). Sample size, test length, and skewness of item response distributions were manipulated in this simulation study. Generally speaking, IRT-GRM performed the best and was most robust against skewness. FA-PR and FA-PL performed equally well across almost all conditions but were competitive with IRT-GRM only when item responses were normally distributed. SSI practice might be slightly worse than the two FA procedures when item responses were normally distributed, but it was better than them when item responses were highly skewed. WMDU performed as well as did SSI only when item responses were normally distributed or moderately skewed and sample size was large for MDS models (e.g., N=100). IMDU and EMDU were even worse than WMDU and appeared not appropriate for Likert-type data.

Data collected in social science frequently produces ordinal variables. However, educational and psychological researchers are frequently interested in some continuous latent traits behind those ordinal data. Likert-type scales are probably one of the most popular ordinal scales which try to measure a continuous latent trait.

There are many statistical ways of modeling and estimating a latent variable. First, classical true score test theory decomposes the observed variable into two latent parts, i.e., the true- and the error-score variables (Lord & Novick, 1967). The true score is defined as the expected value of the observed score or is regarded as the observed score of a person on a homogeneous test of infinite length. Under this conception, the observed score is taken as the unbiased estimator of the true score. Given Likert-type items, however, many researchers are reluctant to assign successive integers to the item response categories arbitrarily and then simply sum up the item scores to obtain the observed score as the estimator of the true score. They are unwilling to assume that an interval scale exists in the Likert-type items. For simplicity, however, this SSI (Sum of Successive Integers) approach is frequently employed by research practioners. This practice is also justified by the results of some Monte Carlo studies (e.g., Bollen & Barb, 1981; Jenkins & Taber, 1977; Lissitz & Green, 1975), which showed that errors caused by regarding ordinal scales as interval ones were small enough to be ignored as long as a minimum of 5 response categories were employed. Recently, Cohen (1990) still suggested that unit weights rather than optimal weights from regression or factor analysis be assigned to the individual Likert-items to be combined. Because the SSI approach is so widely employed, it was served in the current study as the baseline for comparing the other latent variable models.

The second way to identify one or more latent variables in psychology is through the statistical concept of local independence, which underlies a body of latent variable models including factor analysis (FA), item response theory (IRT), latent profile analysis and latent class analysis (Bartholomew, 1987). Because the last two analyses assume a non-metric latent variable, they were not included in the current study.

A psychological interpretation of the term "local independence" is that an individual's systematic performance can be completely explained by some underlying traits so that, given his values on these traits, no more information can be learned from him/her. In other words, his/her performance should be random once his/her values on these traits have been

fixed (Anderson, 1959). Translated into statistical terms, local independence is the statistical independence of residuals once the latent variables in the model constitute the complete latent space and their values are fixed.

However, statistical independence may be defined in different ways in different models. FA defines it in terms of correlations or covariances in the population. IRT defines it in terms of probabilities. More specifically, FA tries to derive some common factors which constitute the complete latent space that can "explain away" the observed pairwise correlations between manifest variables. IRT requires the assumption of unidimensionality and utilizes its consequence of local independence to estimate person and item parameters in such a manner that "the probability of the response pattern for each examinee is equal to the product of the probability associated with the examinee response to each item" at a fixed value of the latent trait (Hambleton & Swaminathan, 1985, p.23).

The third way to define a latent variable is in terms of spatial distance. Typical examples of spatial models can be found in the family of multidimensional scaling (MDS). Like common FA, MDS attempts to search for a small number of latent variables which can explain the observed relationships among a much larger number of stimuli (Loehlin, 1987). In other words, it tries to construct a spatial configuration which can represent the quantity or rank order information of all observed pairwise proximities. Unlike FA, however, MDS does not utilize the principle of local independence. Instead, it (especially the classical metric MDS) employs algorithms similar to the ones applied by principal components analysis (Davison, 1985; Davison & Srichantra, 1988; Rodgers & Young, 1981).

Because FA, IRT, and MDS models utilize sophisticated procedures for recovering metric information from non-metric data, they will be explored further in the following sections.

FA

The basic theories of FA can be summarized in matrix form. Let x represent the vector of manifest variables, W the matrix of factor loadings, f the vector of common factor scores, and e the vector of residuals, then

$$x = W f + e. \qquad (1)$$

The covariance matrix Σ of x is then given by

$$\Sigma = W \Phi W' + \Theta, \qquad (2)$$

where Φ is the covariance matrix of f, and s is the vector of unique variances of x. If the distributions of the residual e are assumed to be multivariate normal, then the conditional distribution of x will also be multivariate normal. In addition, if latent factors, fs, follow multivariate normal distributions, then the marginal distributions of the manifest variables, xs, can be derived and will also be multivariate normal. This marginalization process is usually employed to facilitate the estimation of latent variables given discrete data (Mislevy, 1986).

Given ordered polychotomous indicators, the question arises as to whether or not it is appropriate to compute Pearson correlations for ordinal-level variables and to perform FA on a Pearson correlation matrix (FA-PR). Labovitz (1967, 1970) provided very influential justification for using Pearson correlations for ordinal variables. In his simulation studies, he demonstrated that "monotonic random scoring" systems, which indicated randomly stretched scales, are highly related with "equal distance scoring" systems. He argued that the errors due to treatment of ordinal variables as interval were small enough to be ignored. His studies, however, were limited to the following conditions: a) The underlying latent variable was uniformly distributed; and b) As many as thirty-one categories of the ordinal variable were used. O'Brien (1979) found that the Pearson correlations between stretched scales and the equal distance scoring system were high and increased with the number of categories only when the underlying variable was normally or uniformly distributed and the number of categories was greater than four. Bollen and Barb (1981) also found that it was appropriate to compute Pearson r's on ordinal variables obtained from classifying continuous variables when: a) the two continuous variables were normally distributed, b) they did not correlated highly, and c) at least five categories were used for collapsing.

Given discrete data, it is well known that the Pearsonian correlation is not free to range from -1 to 1 when the two correlated variables are skewed highly in opposite directions (Carroll, 1961; Muthén, 1983). In a simulation study, Olsson (1979a) indeed found that the maximum likelihood FA may create a substantial lack of fit of the true model when it was performed on the Pearsonian correlations computed from successive integers assigned to ordinal categories. His findings were especially true when the observed variables were skewed in opposite directions and the true loadings were high. He also found that the classification of continuous scores into categorical scores attenuated the estimates of factor loadings. The attenuation increased when the variation in skewness of manifest variables increased and the number of scale steps

decreased. Given these results, Olsson suggested that researchers perform a FA on polychoric correlations (FA-PL) when observed variables were obtained from a classification of some continuous latent variables. This suggestion was also made by some other researchers (e.g., Carroll, 1961; Muthén, 1983; Muthén, 1984).

The polychoric correlation coefficient is a generalization of the tetrachoric correlation coefficient to the polychoric case. It estimates the relationships between two latent continuous variables both of which are assumed to be normally distributed and are measured by ordinal scales. Olsson (1979b) presented two ML estimation procedures for the polychoric case. The first procedure estimates the threshold values and the polychoric correlation simultaneously. The second procedure has two steps: a) to estimate threshold values with the inverse of the normal distribution function evaluated at the cumulative marginal proportions of each variable, and b) to obtain the ML estimate of the polychoric correlation, given the threshold values. The two-step procedure has been employed in the LISREL program (Jöreskog & Sörbom, 1984, 1989).

Jöreskog and Sörbom (1988) showed in a simulation study that: a) polychoric correlations were not sensitive to the marginal distributions of the observed variables; b) compared to Spearman's rank correlations, Kendall's tau-b correlations, and product-moment correlations, polychoric correlations were the best estimators of the true latent relationships; and c) polychoric correlations appear to be the only consistent estimators of the true latent relationships. In another simulation study, Babakus, Ferguson and Jöreskog (1987) also found that, compared to product-moment, Spearman's rho, and Kendall's tau-b correlations, polychoric correlations gave the most accurate estimates of the true latent correlations and factor loadings but produced the worst goodness-of-fit values.

Given the above results, the current study predicted that FA-PL should have a better performance than FA-PR in terms of the accuracy of estimating the underlying trait from Likert-type items. This prediction should be especially true when item responses were non-normally distributed.

IRT

IRT typically has been developed for scaling dichotomous data onto an equal-interval scale. For dealing with ordinal polychotomous data, Samejima (1969) developed a two-stage procedure to derive the probability of an individual selecting a particular response category in an polychotomous item. In the first stage, an item with response categroy 0, 1, ..., m was viewed as a combination of m dichotomous items and the two-

parameter model was applied to model the cumulative probability of an individual responding to a particular or higher category. This idea is expressed by the following equation:

$$\sum_{j=k}^{m} \pi_{nij} = \frac{\exp[a_i(\theta_n - b_{ik})]}{1 + \exp[a_i(\theta_n - b_{ik})]}, \quad k = 1, \ldots, m \qquad (3)$$

where π_{nij} is the probability of person n responding to item i with response category j, a_i is the discrimination parameter for item i, b_{ik} is the category boundary between response category k and k-1 of item i, and θ_n is the latent trait level of individual n. Note that the probability of responding to category 0 or higher covers the complete probability space and was set to be one, that is,

$$\sum_{j=0}^{m} \pi_{nij} = 1. \qquad (4)$$

In the second stage of Samejima's procedure, the probability of an individual responding to a particular response category j is given by:

$$\pi_{nij} = \sum_{j=k}^{m} \pi_{nij} - \sum_{j=k+1}^{m} \pi_{nij}, \quad k = 0, \ldots, m. \qquad (5)$$

Equation 5 is the general form for drawing the operating characteristic curves for a graded response item. From equation 3, 4, and 5, it can be derived that the probability of responding to the first and the last category is given by:

$$\pi_{ni0} = 1 - \sum_{j=1}^{m} \pi_{nij}, \qquad (6)$$

$$\text{and } \pi_{nim} = \frac{\exp[a_i(\theta_n - b_{im})]}{1 + \exp[a_i(\theta_n - b_{im})]} \qquad (7)$$

respectively. Equation 7 is equivalent to equation 3 when k=m.

Samejima's graded response model (IRT-GRM) allows a parameter for each item to have a different discrimination power. In addition, it allows the distance between response scale thresholds to vary across items. This model is conceptually similar to the FA-PL model. The applicability of the IRT-GRM to Likert-type scales has been demonstrated by

several studies (e.g., Dodd, 1984; Dodd, Koch & DeAyala, 1988; Koch, 1983; Thissen & Steinberg, 1988).

MDS

The basic model of non-metric MDS is as follows:

$$\delta_{ij} \approx f(d_{ij}) = f\{[\Sigma_k(x_{ik} - x_{jk})^2]^{1/2}\}, \quad (8)$$

where f is a monotone function such that

$$d_{ij} < d_{i'j'} \Rightarrow f(d_{ij}) < f(d_{i'j'}). \quad (9)$$

The basic notion underlying the nonmetric MDS is that sufficient nonmetric constraints can act like metric constraints (Shepard, 1962). Given a dimensionality and a sufficient number of points, the solution configuration that complies with the set of inequalities becomes highly constrained (MacCallum, 1974). Therefore, it is possible to recover the metric information from nonmetric data.

The distance models for preference data are often called unfolding models. Usually, unfolding models attempt to model the proximities between the subjects and the stimuli (e.g., items) instead of the proximities between the pairs of stimuli. Given one rectangular matrix of preference data, the following two classical unfolding models are most probably considered: Coombs' (1964) internal unfolding (IMDU) and Carroll's (1972) external unfolding (EMDU) procedures. Coombs' model has the following form:

$$\delta_{is} \approx f_s(d_{is}) = f_s\{[\Sigma_k(x_{ik} - x_{sk})^2]^{1/2}\}, \quad (10)$$

where δ_{is} refers to the degree of subject s' preference for stimulus i, f_s is the monotonic function specific to subject s, d_{is} is the estimated distance between the stimulus i and the subject s in the model, x_{ik} represents the location of stimulus i along dimension k, and x_{sk} represents the ideal point of subject s along dimension k. Evidently, this is a "point model" instead of a "vector model." It is assumed that the more the subject's ideal point is close to the location of the stimulus' scale value, the higher the subject's preference will be, regardless of whether the ideal point is at the positive or the negative side of the location of the scale value on the dimension.

Carroll's (1972) EMDU models have exactly the same assumptions except that the stimulus/item coordinates are assumed to be known from the theory or from a previous estimation. The EMDU procedures may include the following

steps: a) calculating dissimilarities among all pairs of items using squared Euclidean distances; b) subjecting the matrix of dissimilarity data to the classical MDS procedure for estimating item coordinates; and c) fixing the item coordinates and obtaining estimates of person parameters with unfolding models. It should be obvious that EMDU, like FA and IRT models, employs a two-stage procedure--namely, it first estimates the item parameters and then estimates the person parameters assuming that the item parameters are known.

The two classical MDU models may be applicable to the test data where the trace line of each item is a cumulative function, i.e., the relationships between the latent trait and responses to each item are monotonic. Some MDS theorists have argued that monotonic relationships between vectors are able to be modeled by positioning subjects' ideal points far away from the location of the items' ideal points in the direction of the latent dimension (Schiffman, Reynolds & Young, 1981, p.262). In other words, clusters of persons and items may be formed separately in opposite directions of the solution space. In this situation, the solution may look "degenerate" but the ideal points of the subjects and items may still be estimated appropriately. In a simulation study, Fitzpatrick (1989) generated data according to the Rasch model, which assumes a monotonic relationship between the latent trait and the probability of a correct response to an item. He found that the two classical MDU procedures often produced what many researchers call "degenerate" solutions with high stress, but the recovered person and item coordinates still correlated highly with the true parameters. His findings were nevertheless restricted to the situation where dichotomous items with equal discrimination parameters were used. Dichotomous items might be adequately represented by points because only one threshold value is involved for each item. In addition, the assumption of equal discrimination of items fits the classical MDU approaches, which explicitly model the differences between difficulty parameters rather than discrimination parameters. It is not clear how classical MDU approaches can model the differences among item difficulty levels and among item discrimination parameters simultaneously.

For analyzing Likert-type data, the two classical MDU procedures may be especially inappropriate. Every Likert-type item with n response categories can be decomposed into n-1 dichotomous items which attempt to measure a wide range of the latent trait. The two classical MDU models will simply compress each Likert item into one geographical point and bypass the estimation of threshold values and of discrimination parameters. In theory, Coombs (1964) and Davison (1983) tended to argue against, while Coombs and Smith (1973), Lingoes (1972), and Young and Lewyckyj (1979) tended to argue for, the

application of the classical MDU models to profile data. In the published literature, few applied studies can be found. In an empirical study where the IMDU model was applied to two Likert format instruments, Koch (1984) found that degenerate solutions were paramount regardless of what computer programs was used, whether the row conditionality or the matrix conditionality assumption was made, and whatever number of dimensions was used. In the MDS field, therefore, it seems desirable to search for a better alternative to the classical MDU models for the analysis of Likert-type data.

It is clear that both FA with polychoric correlations and the graded response model in IRT are able to estimate not only the item threshold values but also the relationship between each item and the latent dimension. Therefore, one promising direction for the MDU models to analyze Likert-type data is to regard each response category of each item as one stimulus and to take the relationship between each item and the latent dimension as the weight of each item on the latent dimension. This situation calls for a weighted unfolding model (WMDU) (Young & Lewyckyj, 1979; Young, 1984).

The WMDU model can be viewed as the individual differences scaling model (Carroll & Chang, 1970) extended to preference data. For simplicity of expression, the model is shown with matrix algebra as follows:

$$\delta_{ijk} \approx f_i(d_{ijk}) = f_i\{[(\underline{y}_i - \underline{x}_j)'W_k(\underline{y}_i - \underline{x}_j)]^{1/2}\}, \quad (11)$$

where δ_{ijk} is the proximity between row i and column j for matrix k, f_i is the monotonic function for row i, d_{ijk} is the estimated distance corresponding to δ_{ijk}, $\underline{y}_i$ is the vector for row i, $\underline{x}_j$ is the vector for column j, and the diagonal W_k is the weight for matrix k. In the current situation, subscripts i, j, and k correspond to subjects, response categories, and items. It can be seen that the weight for "individual difference" is applied to model the differences in item discriminations or factor loadings on the latent dimension.

To apply WMDU to Likert-type data, the assumption that the response categories of each item are in order must hold. Consequently, if a person i chooses response category j and j is the lowest category, then his/her preference about the response categories can be inferred to be: j, j+1, j+2, ..., to the highest category. If j is the highest category, then his/her preference order should be: j, j-1, j-2, ..., to the lowest category. If j is the middle category, then his/her preference order is: j, j±1, j±2, ..., to the lowest or the highest category. If this data transformation were done for each item and each person, then a "three-way three-mode" data set would be obtained. This data set would contain multiple

matrices of rectangular data with rows corresponding to subjects, columns corresponding to response categories, and matrices corresponding to items. If WMDU were applied to this kind of data, coordinates of subjects and of item response categories and the weight for each item can be estimated for the latent dimension. Although only one set of coordinates of response categories is estimated for all items, this model preserves each Likert-type item as a vector and, therefore, is able to model the relationship between each item and the latent dimension. It was predicted that the WMDU model should outperform the two classical unfolding models.

FA vs. IRT vs. MDS

The FA with either the two-stage maximum-likelihood estimators of polychoric correlations or Pearsonian correlations is a limited information approach, but the marginal maximum likelihood (MML) estimation for the IRT-GRM is a full information approach. The limited-information approach utilizes only the information in the lower order margins in the full k^p (p variables with k response categories) contingency table while the full-information approach utilizes all information in the table (Mislevy, 1986). The matrix of second-order cross-products analyzed by FA is a sufficient statistic only when all latent and manifest variables are approximately normally distributed (Mooijaart, 1985). In this situation, the advantages of the full information approaches of estimation may disappear. Therefore, the performance of FA-PR and FA-PL are predicted to be as good as IRT-GRM when item responses were normally distributed and to be worse than it when distributions of item responses were skewed. These predictions are also based on the consideration that the current study investigates only the unidimensional case. Mislevy (1986) pointed out that the MML estimation method based on the full information is preferable for long tests with few factors.

Most empirical studies which have compared MDS with FA models typically have concentrated on the number of dimensions, the congruence of dimensions, and the interpretability of the constructs that emerged from the analyses (e.g., Dancer, 1985; Gillespie, 1989; Koch, 1984). Most of them ignored MDU and individual differences MDS and, therefore, ignored the person parameters. Based on the facts that the WMDU model estimates one set of response-category coordinates for all items while FA-PL estimates different threshold values for different items, it was predicted that the WMDU model should perform slightly worse than FA-PL. These relative advantages of FA-PL to WMDU, however, may not be evident when all distributions of item responses are approximately normal, where response categories had similar function across items. Regarding comparison with FA-PR, WMDU was predicted to perform equally well. Finally, the

two classical unfolding models, IMDU and EMDU, should perform worse than WMDU, FA-PR and FA-PL because they estimate neither threshold values nor discrimination parameters of items.

As to the comparison of IRT-GRM with the three unfolding models, it was predicted that the WMDU model should perform worse than IRT-GRM because WMDU gives only one set of estimates of threshold values for all items, while the IRT-GRM gives different sets of estimates of threshold values for different items and therefore is more flexible. These relative advantages of IRT-GRM to WMDU may not be evident when all distributions of item responses were approximately normal.

Methodology

The Simulated Situation

For each item, a continuous response variable (Y^*) was generated according to the following formula:

$$Y^* = \beta X + \sqrt{(1-\beta^2)}\, E, \qquad (12)$$

where X was the set of standardized Z-scores on the latent variable, β was a regression weight conceptually corresponding to the relationship between the item and the latent variable, and E was the error or residual vector. Both X and E were generated according to the normal distribution and had a mean of zero and a variance of one. Therefore, Y^* also had a mean of zero and a variance of one. The values of β were systematically chosen and randomly assigned to each item because each item was systematically chosen to have a certain skewness value, which will be explicated later.

Secondly, the integer response score (Y) for each simulated subject was decided according to the following scheme:

$$\begin{aligned} &Y = 1, \text{ if } && Y^* < t_1 \\ &Y = 2, \text{ if } t_1 \le && Y^* < t_2 \\ &Y = 3, \text{ if } t_2 \le && Y^* < t_3 \\ &Y = 4, \text{ if } t_3 \le && Y^* < t_4 \\ &Y = 5, \text{ if } t_4 \le && Y^* \end{aligned} \qquad (13)$$

Note that t1, t2, t3, and t4 were threshold values, which were manipulated in order to obtain desired distributional characteristics of item responses.

Background Conditions

The following conditions were held constant across experimental conditions: 1) The observed variables were ordinal

and had five discrete response categories; 2) There was only one latent dimension underlying the observed variables; 3) The latent variable was continuous; 4) The observed and the latent variable had a monotonic non-decreasing relationship; 5) The latent variable and the measurement error were normally distributed; and 6) The response data were interpersonally comparable (c.f., Brady, 1989), that is, the function which related the latent variable to the response variable was the same for all subjects.

Dependent Variables

Two dependent variables of interest were: 1) the absolute value of Pearson correlations between the recovered and the true person parameters, and 2) the root mean squared differences (RMSD) between the recovered and the true person parameters.

The scaled latent trait continuum had been standardized to have a mean of 0 and a standard deviation of 1 before the two dependent variables were computed. When necessary, it also had been reversed to have the same direction as the true latent trait continuum before it was standardized. This linear transformation was employed in order to eliminate artificial scaling factors in different computer programs while keep the distributional shape of the estimated latent trait continuum unchanged.

Independent Variables

There were four independent variables: 1) Sample sizes (30 vs 100 vs 1000 subjects); 2) Test lengths (12 vs 24 items); 3) Distributional characteristics of item responses (normally distributed vs moderately skewed vs highly skewed vs differentially skewed); and 4) Seven statistical procedures used to recover the true parameters (FA-PL vs FA-PR vs IRT-GRM vs WMDU vs IMDU vs EMDU vs SSI). The first three independent variables were used to form 24 (= 4 X 3 X 2) experimental conditions, within each of which five replications of the simulated data were generated.

Because a sample size of 1000 subjects (Case I) was too large for most current MDS procedures, only FA-PR, FA-PL, IRT-GRM, and SSI procedures were compared with each other in this case. A sample size of 100 subjects (Case II) was very special because all seven statistical procedures were applicable in this case. This size of sample might be insufficiently large for FA and IRT but was considered very large for MDS procedures. Given this case, a comparison of FA's with IRT's robustness against small sample size was available. Given 12 or 24 items, a sample size of 30 subjects (Case III) was evidently

too small for FA and IRT, so that only WMDU, IMDU, EMDU, and SSI were compared with each other.

Insert Table 1 about here.

The distributions of item responses are a function of item threshold values. In terms of standardized Z-scores, the population threshold values were systematically chosen so that each item had a distribution with desired skewness value in the population (see Table 1). Skewness were computed through the following procedures: 1) Successive integers, 1 thru 5, were assigned to the five response categories; 2) Population means and standard deviations of these response scores for each item were obtained with the following two formulas:

$$\mu = E(X) = \Sigma P_i X_i, \quad (14)$$

$$\text{and } \sigma = E(X^2) - \mu^2, \quad (15)$$

where X_i was the set of integer scores ranging from 1 to 5 and P_i was the probability of each integer score; 3) The assigned integer scores were standardized with the obtained mean and standard deviation; 4) Skewness was computed as the third moment of the standardized scores.

Insert Table 2 about here.

Skewness values were deliberately chosen to produce normal, moderately skewed, highly skewed, and differentially skewed distributions of item responses (see Table 1 and Table 2). Within each distributional condition, twelve β weights (.35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90) were randomly assigned to the 12 items. The twelve items serve as "core items" and were duplicated in order to obtain the condition of 24 items.

The expected proportion of subjects responding in each response category was computed as the area between threshold values under the normal curve. An effort was made to avoid the expected number of cases in each response category being zero when extreme threshold values were to be selected.

Programs and Procedures for Estimating

For applying the FA-PR and the FA-PL procedures, the LISREL VI computer program (Jöreskog & Sörbom, 1984) was employed. In general, ML estimation was adopted. However, when

the matrix of polychoric correlations was not positive definite, the unweighted least square estimation was used. Starting values for all parameters were set at 0.5.

For applying the IRT-GRM procedure, the MULTILOG computer program (Thissen, 1986) was employed. ML estimation was used to estimate item while conditional marginal ML was used to estimate person parameters. By default, the starting value for descrimination parameter was 1.0, while the starting values for the threshold parameters were -1.39, -0.405, 0.405, and 1.39 respectively.

For applying the WMDU, IMDU, and the EMDU procedures, the SAS PROC ALSCAL computer program (Young & Lewyckyj, 1979) was employed. Because Young and Lewyckyj suggested that results of the analysis might be less valid with similarities data, all similarities data were converted into disimilarities data.

A scheme of recoding raw data was required before the WMDU model could be applied. Let Y represent the integer response score, while C_1, C_2, C_3, C_4, and C_5 represents the five response categories as column stimuli. According to the following scheme, data recoding was made for each subject:

$$\begin{array}{l} \text{If } Y=1, \text{ then } C_1=1,\ C_2=2,\ C_3=3,\ C_4=4,\ C_5=5; \\ \text{Else if } Y=2, \text{ then } C_1=2,\ C_2=1,\ C_3=2,\ C_4=3,\ C_5=4; \\ \text{Else if } Y=3, \text{ then } C_1=3,\ C_2=2,\ C_3=1,\ C_4=2,\ C_5=3; \\ \text{Else if } Y=4, \text{ then } C_1=4,\ C_2=3,\ C_3=2,\ C_4=1,\ C_5=2; \\ \text{Else if } Y=5, \text{ then } C_1=5,\ C_2=4,\ C_3=3,\ C_4=2,\ C_5=1. \end{array} \qquad (16)$$

After the above recoding procedure was applied to every subject, multiple rectangular matrices were formed with row stimuli corresponding to subjects, column stimuli corresponding to response categories, and matrices corresponding to items. Data in each rectangular matrix were considered to be ordinal and row-conditional. Tied data were set to be untied. Number of dimensions was set at two, which was the minimum number required by the individual differences MDS. The subject coordinates on the first dimension were taken as estimates of the latent trait.

The usual subject-by-item rectangular data were directly submitted to the ALSCAL program for IMDU analysis. The data were considered to be ordinal, and matrix-conditional. Tied data were set to be untied. Number of dimensions was set to be one. Euclidean model was assumed.

Three steps of procedure were needed for EMDU analysis. First, for transforming rectangular data into symmetric proximity data, Euclidean distances between all possible pairs of items were computed as the sum (over subjects) of the

squared differences between the paired scores. Second, the symmetric proximity data was submitted to the ALSCAL program and analyzed by the classical non-metric MDS model. The data were considered to be ordinal and matrix-conditional. Tied data were set to be untied. Number of dimensions was set to be one. Euclidean model was assumed. Third, the IMDU procedure described in the last paragraph was again employed except that item coordinates estimated from the second step were read in and fixed for estimating subject coordinates. Note that the fix option fixed the relative positions of items but fixed neither the origin of the stimulus space nor the lengths of dimensions (Young & Lewyckyj, 1979).

Results

Descriptive ANOVA was used for analysis. Given each experimental condition, the mean and standard deviations of five correlations/RMSDs from five replicated samples were presented in Table 3 and Table 4 for Case I, in Table 5 and Table 6 for Case II, and in Table 7 and Table 8 for Case III. The patterns of results were highly similar acorss the two assessing criteria, i.e., correlations and RMSDs. Therefore, the following findings may be regarded as valid across both criteria.

Case I: N=1000

In this case, IRT-GRM, FA-PL, FA-PR, and SSI were compared with one another. Some interesting results could be derived from Table 3 and Table 4.

Insert Table 3 and Table 4 about here.

---

1) Generally speaking, IRT-GRM performed best among the four procedures in terms of recovering the latent trait parameters. However, a few interactions between the recovering procedures and the distributions of item responses could also be observed. First, compared to FA-PL and FA-PR, IRT-GRM performed just as well as did the two FA procedures when the item responses were normally distributed but it tended to outperform them when the distributions of item responses became skewed. The advantage of IRT-GRM was especially evident when the distributions of item responses were highly skewed. Second, compared to the common SSI procedure, IRT-GRM performed slightly better when the distributions of item responses were normal, moderately skewed, or differentially skewed but much better when the distributions were highly skewed. It seemed

that IRT-GRM was more robust against skewness than the other three procedures involved. These results were true across the conditions of test length.

2) FA-PR performed as well as did FA-PL. This finding was true across all experimental conditions of test length and of response distributions.

3) Both FA-PL and FA-PR were only slightly better than SSI when item responses were normally distributed. This advantage of factor analysis became even smaller when distributions of item responses were moderately or differentially skewed. Moreover, both FA-PL and FA-PR performed slightly worse than SSI when the distributions of item responses were highly skewed. It seemed that factor analysis was more sensitive to the disturbance of skewness than was SSI. These results were true across the conditions of test length.

4) It is not surprising that both test length and distributions of item responses have a main effect. All four procedures performed better in the condition of 24 items than in the condition of 12 items. They performed best when the distributions of item responses were normal, next best when the distributions were moderately or differentially skewed, and worst when the distributions were highly skewed.

5) All recovering procedures were pretty stable across replicated samples according to the fact that all standard deviations of the five correlations/RMSDs were less than or equal to 0.01.

Case II: N=100

In this case, all seven procedures were compared with one another. The following results were derived from Table 5 and Table 6.

Insert Table 5 and Table 6 about here.

---

1) Generally speaking, IRT-GRM performed best among the seven procedures in terms of recovering the latent trait parameters. This conclusion depended somewhat on the condition of item response distributions. When the item responses were normally distributed, IRT-GRM performed just as well as did FA-PR, FA-PL, WMDU, and SSI but much better than did EMDU and IMDU. However, it outperformed all other six procedures when the distributions of item responses were moderately skewed or highly skewed. When the distributions of item responses were

differentially skewed, IRT-GRM performed slightly better than did FA-PR, FA-PL and SSI. It seemed that IRT-GRM was more robust against skewness than were the other six procedures. These results were even clearer when number of items was larger and when RMSDs instead of correlations were assessed.

2) Generally speaking, FA-PR, FA-PL and SSI had similar performances across most conditions. However, a slight tendency might be observed: although FA-PR and FA-PL performed as well as or even slightly better than did SSI when the distributions of item responses were normal or moderately skewed, they performed worse than did SSI when the distributions were highly skewed. These results implied that both FA-PL and FA-PR might be more severely disturbed by a high degree of skewness than was SSI.

3) WMDU performed as well as did IRT-GRM, FA-PR, FA-PL, and SSI when distributions of item responses were normal. It performed worse than IRT-GRM when distributions of item responses were moderately or highly skewed. It performed much worse than IRT-GRM, FA-PR, FA-PL, and SSI when item responses were differentially skewed.

4) Generally, the IMDU procedure performed worse than the traditional SSI procedure but better than EMDU in almost all conditions. It performed worse than did WMDU in most conditions except when item response distributions were differentially or highly skewed. If results from Case III (in Table 7 and Table 8) were also taken into consideration, it became clear that WMDU was generally better than IMDU.

5) EMDU performed worst among the seven procedures in almost all conditions.

6) With the exception of the two classical unfolding procedures (IMDU and EMDU), all procedures performed best when item responses were normally distributed and worst when distributions were highly skewed.

7) Regardless of conditions of item response distributions, increasing test length improved estimation for all procedures except the two classical unfolding models.

8) In most conditions, the two classical unfolding models were less stable than the other five procedures because they occasionally produced inappropriate solutions which completely failed to recover the true person parameters.

Case III: N=30.

In this case, WMDU, IMDU, EMDU and SSI were compared with one another. The following results could be derived from Table 7 and Table 8.

Insert Table 7 and Table 8 about here.

---

1) Generally speaking, the common SSI procedure performed as well as or better than all three unfolding procedures in all conditions of test length and response distributions.

2) WMDU performed better than the two classical unfolding procedures, EMDU and IMDU, in all conditions.

3) IMDU performed better than did EMDU when the distributions of item responses were normal or differentially skewed. EMDU outperformed IMDU when distributions of item responses were highly skewed. The relative merits of IMDU and EMDU were not clear when distributions of item responses were moderately skewed.

4) Both of the SSI and the WMDU procedures had more stable performances across replications than did the two classical unfolding procedures. The latter two procedures were unstable because they sometimes produced inappropriate solutions which completely failed to recover the true person parameters of a few samples.

5) Both SSI and WMDU performed best when item responses were normally distributed and worst when item responses were highly skewed. However, SSI performed better when item responses were differentially skewed than when item responses were moderately skewed, while WMDU had the opposite pattern of performances.

6) The above outcomes became clearer when number of items became larger.

Conclusions across Cases

1) From the best to the worst, the general performances of the seven procedures could be ordered as follows: a) IRT-GRM, b) FA-PL and FA-PR, c) SSI, d) WMDU, and e) IMDU and EMDU.

2) IRT-GRM was more robust against skewness than were FA-PL, FA-PR, and SSI even when sample size was as small as N=100.

3) FA-PL and FA-PR were competitive with IRT-GRM only when item responses were normally distributed.

4) FA-PL and FA-PR performed equally well across almost all conditions.

5) The common SSI practice might be slightly worse than the two FA procedures when item responses were normally distributed, but it was better than them when item responses were highly skewed.

6) The WMDU procedure was generally not a better alternative to the common SSI practice. It performed as well as did SSI only when item responses were normally distributed or moderately skewed and when sample size was large for MDU (e.g., N=100).

7) IMDU and EMDU were even worse than WMDU and should be avoided for Likert-type data. The reader should recall that the IMDU and EMDU procedures modeled the data in a completely different way than did the WMDU procedure.

Discussion

FA-PR vs. FA-PL

It was predicted that FA-PL should perform better than FA-PR. This prediction was made because previous literature had shown that FA-PL gave more accurate estimates of the pairwise true latent correlations and of the factor loadings than did FA-PR. This prediction, however, was not confirmed by the current results. When the recovered correlations and the estimated factor loadings were examined, Pearson correlations rather than polychoric correlations were indeed found to systematically underestimate the true latent correlations. It was also found that FA-PR rather than FA-PL systematically underestimated the true factor loadings. In addition, these underestimations were especially obvious when item responses were highly or differentially skewed. All of these findings were consistent with results of previous studies. Nonetheless, the current results showed that the systematic underestimation of both true latent correlations and factor loadings did not hamper the estimation of the true latent trait. Across almost all conditions set by the current study, FA-PR estimated the latent trait as accurately as did FA-PL. It seemed that the systematic underestimation of factor loadings was compensated by systematic overestimation of regression weights for factor scores when the latent trait was to be estimated. This compensation phenomenon was even found for dichotomous items in a later study (Chan, in preparation).

WMDU vs. IMDU vs. EMDU

As predicted, the WMDU model were found to outperform the two classical unfolding models in almost all experimental conditions. This result implied that, if procedures in the MDS family are to be employed for analyzing Likert-type data, the WMDU model should be a more reasonable alternative to the two classical unfolding models.

IMDU and EMDU were expected to completely fail in the situation where all item responses were normally distributed and to perform best when item responses were differentially skewed. This prediction was not confirmed. The performances of IMDU and EMDU were very unstable--they fluctuated from one condition to another. The underlying mechanism of the two classical unfolding procedures to model the monotonic relationships between the Likert-type items and the latent trait seems puzzling. The current study suggested that IMDU and EMDU should not be applied to Likert-type items.

IRT vs. FA vs. MDS

Conceptually, IRT-GRM, FA-PL, FA-PR, and WMDU are all able to model both discrimination parameters and threshold values in Likert-type items. Their predicted relative merits were generally confirmed. IRT's non-linear formulation of the relationship between responses and latent traits seemed to have some advantages, especially when distributions of item responses were skewed. The current study suggests that when distributions of item responses are moderately or highly skewed, IRT-GRM is the favored choice for estimating the latent trait. When item responses are normally distributed, IRT-GRM, FA-PL, FA-PR, and WMDU are all reasonable choices, which may be decided according to other objectives.

SSI vs. Other Procedures

The common SSI practice for scaling the Likert-type items has usually been criticized for its strong assumption of equal intervals in the ordinal scales. The present study found that this strong assumption was not very harmful given that the number of response categories for each item was five. In the present study, the SSI procedure might be slightly worse than the IRT-GRM but was comparable with the two sophisticated FA procedures. The SSI procedure was even more robust against skewness than the two FA procedures. In addition, it was as good as or better than the WMDU and was definitely better than the two classical unfolding procedures. If computer time and computational simplicity were also taken into consideration,

the SSI procedure became more attractive than the other six procedures.

Limitations

Regarding the latent space, the current study investigated only the unidimensional case. Some of the current findings may not be generalizable to multidimensional cases. For example, IRT-GRM which utilizes MML estimation based on the full information approach is preferable for long tests with few factors. However, the extention of IRT-GRM to multidimensional case is still under developing.

Although Likert-type items are often based on 5-point response scales, general attitude measurements were also frequently implemented with non 5-point scales. It is unknown whether or not the current findings were generalizable to non 5-point scales. Although the current study had no reason opposing this kind of generalization, a further study is needed for providing some evidence. It is possible that, due to the increasing or decreasing number of parameters to be estimated, the relative merits of the seven statistical procedures may change.

There are many kinds of item-response distributions which are qualitatively different from one another. For example, U-shaped, uniform, or other irregular distributions are occasionally encountered in practice. The current study, however, investigated only normal and skewed distributions. It is still unknown how accurate the seven statistical procedures would be in estimating the latent trait from Likert-type data given those response distributions which were not investigated.

References

Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In Probability and Statistics, The Harald Cramér Volume, ed. U. Grenander, Stockholm: Almqvist & Wicksell and Wiley, 9-38.

Babakus, E., Ferguson, C. E., Jr. & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. Journal of Marketing Research, 24, 222-228.

Bartholomew, D. J. (1987). Latent variable models and factor analysis. New York: Oxford University Press.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. American Sociological Review, 46, 232-239.

Brady, H. E. (1989). Factor and ideal point analysis for interpersonally incomparable data. Psychometrika, 54, 181-202.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26, 347-371.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 238-319.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), Multidimensional scaling: Theory and applications for the behavioral sciences, Vol. I: Theory. New York: Seminar Press.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Coombs, C. H. (1964). A theory of data. New York: Wiley.

Coombs, C. H., & Smith, J. E. K. (1973). On the detection of structure in attitudes and developmental processes. Psychological Review, 80, 337-351.

Dancer, L. S. (1985). On the multidimensional structure of self-esteem: Facet analysis of Rosenberg's self-esteem scale. In D. Canter (Ed.), Facet Theory: Approaches to Social Research. New York: Springer-Verlag.

Davison, M. L. (1983). Multidimensional scaling. New York: John Wiley & Sons.

Davison, M. L. (1985). Multidimensional scaling versus components analysis of test intercorrelations. Psychological Bulletin, 97, 94-105.

Davison, M. L., & Srichantra, N. (1988). Acquiescence in components analysis and multidimensional scaling of self-rating items. Applied Psychological Measurement, 12, 339-351.

Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models. Doctoral Dissertation, University of Texas at Austin.

Dodd, B. G., Koch, W. R. & DeAyala, R. J. (1988). Computerized adaptive attitude measurement: A comparison of the graded response and rating scale models. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Fitzpatrick, S. (1989). Rasch model parameter estimation: A comparison of IRT and nonmetric multidimensional scaling methods. Paper presented in the annual conference of the American Educational Research Association.

Gillespie, M. K. (1989). A comparison of factor analysis and multidimensional scaling applied to survey data. Dissertation of the University of Texas at Austin, Austin, TX.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory. Boston: Kluwer.

Jenkins, G. D., Jr., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. Journal of Applied Psycholgoy, 62, 392-398.

Jöreskog, K. G., & Sörbom, D. (1984). LISREL VI: User's guide (3rd ed.). Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1988). PRELIS: A preprocessor for LISREL. Mooresville, IN: Scientific Software, Inc.

Jöreskog, K. G., & Sörbom, D. (1989). LISREL 7: User's Reference Guide. Mooresville, IN: Scientific Software, Inc.

Koch, W. R. (1983). Likert-type scaling using the graded response latent trait model. Applied Psychological Measurement, 7, 15-32.

Koch, W. R. (1984). Degenerate solutions in multidimensional unfolding. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Labovitz, S. (1967). Some observations on measurement and statistics. Social Forces, 46, 151-160.

Labovitz, S. (1970). The assignment of numbers to rank order categories. American Sociological Review, 35, 515-524.

Lingoes, J. C. (1972). A general survey of the Guttman-Lingoes nonmetric program series. In R. N. Shepard et al. (Eds.), Multidimensional scaling: Volume I (pp. 49-68). New York: Seminar Press.

Lissitz, R. W. & Green, S. B. (1975). Effect of the number of scale points on reliability: a Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.

Loehlin, J. C. (1987). Latent variable models: An introduction to factor, path, and structural analysis. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading: Addison-Wesley.

MacCallum, R. C. (1974). Relations between factor analysis and multidimensional scaling. Psychological Bulletin, 81, 505-516.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Mooijaart, A. (1985). Factor analysis for non-normal variables. Psychometrika, 50, 323-342.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometries, 22, 43-65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

O'Brien, R. M. (1979). The use of Pearson's r with ordinal data. American Sociological Review, 44, 851-857.

Olsson, U. (1979a). On the robustness of factor analysis against crude classification of the observations. Multivariate Behavioral Research, 14, 485-500.

Olsson, U. (1979b). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44, 443-460.

Rodgers, J. L., & Young, F. W. (1981). Successive unfolding of family preferences. Applied Psychological Measurement, 5, 51-62.

Samejima, F. A. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika (Monograph Supplement No. 17).

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). Introduction to multidimensional scaling. New York: Academic Press.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance. I and II. Psychometrika, 27, 125-140, 219-246.

Thissen, D. (1986). MULTILOG: A user's guide. Mooresville, IN: Scientific Software, Inc.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. Psychological Bulletin, 104, 385-395.

Young, F. W. (1984). The general Euclidean model. In H. G. Law et al. (Eds.), Research methods for multimode data analysis. New York: Praeger. pp. 440-465.

Young, F. W., & Lewyckyj, R. (1979). ALSCAL-4: User's guide. Carrboro, NC: Data Analysis and Theory Associates.

BEST COPY AVAILABLE

Table 1
Population Threshold Values in Terms of Z-Scores

Dist.	Item No.	Threshold Values t1	t2	t3	t4	Skew-ness	β
	1.	-2.076	-.716	.716	2.076	.000	.55
	2.	-2.052	-.712	.712	2.052	.000	.80
	3.	-1.836	-.676	.676	1.836	.011	.35
	4.	-1.812	-.672	.672	1.812	.011	.75
	5.	-1.740	-.660	.660	1.740	.010	.50
(1)	6.	-1.692	-.652	.652	1.692	.000	.45
	7.	-1.668	-.648	.648	1.668	.000	.90
	8.	-1.620	-.640	.640	1.620	.000	.65
	9.	-1.596	-.636	.636	1.596	.000	.40
	10.	-1.548	-.628	.628	1.548	-.010	.85
	11.	-1.500	-.620	.620	1.500	-.010	.70
	12.	-1.476	-.616	.616	1.476	-.010	.60
	1.	-1.645	-.862	-.331	1.100	-.720	.50
	2.	-1.345	-.942	-.415	.900	-.778	.75
	3.	-1.645	-1.002	-.478	.750	-.835	.80
	4.	-1.645	-1.062	-.541	.600	-.878	.65
	5.	-1.645	-1.122	-.604	.450	-.989	.45
(2)	6.	-1.645	-1.142	-.625	.400	-1.021	.90
	7.	-1.645	-1.162	-.646	.350	-1.083	.70
	8.	-1.645	-1.202	-.688	.250	-1.129	.40
	9.	-1.645	-1.222	-.709	.200	-1.172	.85
	10.	-1.645	-1.242	-.730	.150	-1.213	.35
	11.	-1.645	-1.262	-.751	.100	-1.261	.55
	12.	-1.645	-1.282	-.772	.050	-1.307	.60

(continued on next page)

Table 1 (continued)

Dist.	Item No.	Threshold Values t1	t2	t3	t4	Skew-ness	β
	1.	-1.815	-1.316	-.959	-.460	-1.753	.60
	2.	-1.830	-1.336	-.982	-.490	-1.808	.35
	3.	-1.845	-1.356	-1.005	-.520	-1.858	.85
	4.	-1.860	-1.376	-1.028	-.550	-1.916	.50
	5.	-1.875	-1.396	-1.051	-.580	-1.911	.80
(3)	6.	-1.890	-1.416	-1.074	-.610	-2.035	.45
	7.	-1.905	-1.436	-1.097	-.640	-2.095	.70
	8.	-1.920	-1.456	-1.120	-.670	-2.159	.40
	9.	-1.935	-1.476	-1.143	-.700	-2.140	.55
	10.	-1.950	-1.496	-1.166	-.730	-2.282	.75
	11.	-1.965	-1.516	-1.189	-.760	-2.261	.90
	12.	-1.980	-1.536	-1.212	-.790	-2.330	.65
	1.	.790	1.212	1.536	1.980	2.330	.40
	2.	.610	1.074	1.416	1.890	2.035	.75
	3.	.460	.959	1.316	1.815	1.753	.90
	4.	-.050	.772	1.282	1.645	1.307	.55
	5.	-.350	.646	1.162	1.645	1.083	.65
(4)	6.	-1.100	.331	.862	1.645	.720	.45
	7.	-1.645	-.862	-.331	1.110	-.720	.80
	8.	-1.645	-1.162	-.646	.350	-1.083	.50
	9.	-1.645	-1.282	-.772	.050	-1.307	.35
	10.	-1.815	-1.316	-.959	-.460	-1.753	.85
	11.	-1.890	-1.416	-1.074	-.610	-2.035	.60
	12.	-1.980	-1.536	-1.212	-.790	-2.330	.70

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed.

Table 2
Range, Mean, and SD of Skewness Values of 12 Items in Each Condition of Item Response Distributions

Dist.	Range of Skewness	Mean of Skewness	SD of Skewness
(1)	-.010 to .011	.000	.007
(2)	-.720 to -1.307	-1.032	.187
(3)	-1.753 to -2.330	-2.046	.187
(4)	-2.330 to 2.330	.000	1.640

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed.

Table 3
Means and S.D.s of Correlations between True and Recovered Person Parameters for Case I (N=1000)

#Items	Dist.	IRT-GRM	FA-PR	FA-PL	SSI
12	(1)	.95	.95	.95	.93
		(.00)	(.00)	(.00)	(.00)
	(2)	.94	.92	.92	.91
		(.00)	(.00)	(.00)	(.00)
	(3)	.88	.81	.80	.81
		(.00)	(.01)	(.01)	(.01)
	(4)	.93	.93	.93	.91
		(.00)	(.00)	(.01)	(.00)
24	(1)	.98	.98	.97	.96
		(.00)	(.00)	(.00)	(.00)
	(2)	.97	.94	.94	.94
		(.00)	(.00)	(.00)	(.00)
	(3)	.92	.83	.83	.85
		(.00)	(.00)	(.01)	(.01)
	(4)	.97	.96	.96	.96
		(.00)	(.00)	(.00)	(.00)

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed. The number in the parenthesis is the standard deviation of five replicated correlations.

BEST COPY AVAILABLE

Table 4
Means and S.D.s of RMSDs between True and Recovered Person Parameters for Case I (N=1000)

#Items	Dist.	IRT-GRM	FA-PR	FA-PL	SSI
12	(1)	.31	.31	.31	.36
		(.01)	(.01)	(.01)	(.01)
	(2)	.33	.39	.40	.43
		(.01)	(.00)	(.01)	(.01)
	(3)	.49	.62	.62	.61
		(.01)	(.01)	(.01)	(.01)
	(4)	.36	.38	.39	.43
		(.01)	(.01)	(.01)	(.01)
24	(1)	.22	.22	.22	.26
		(.01)	(.01)	(.01)	(.01)
	(2)	.24	.34	.34	.35
		(.01)	(.01)	(.01)	(.00)
	(3)	.39	.58	.59	.55
		(.01)	(.01)	(.01)	(.01)
	(4)	.26	.27	.28	.31
		(.01)	(.01)	(.01)	(.01)

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed. The number in the parenthesis is the standard deviation of five replicated RMSDs.

Table 5
Means and S.D.s of Correlations between True and Recovered Person Parameters for Case II (N=100)

Dist.	IRT-GRM	FA-PR	FA-PL	SSI	WMDU	IMDU	EMDU
				12 items			
(1)	.95	.95	.95	.95	.94	.84	.56
	(.01)	(.01)	(.01)	(.01)	(.01)	(.06)	(.31)
(2)	.95	.92	.92	.91	.92	.86	.76
	(.01)	(.01)	(.01)	(.02)	(.01)	(.03)	(.11)
(3)	.89	.81	.79	.82	.81	.76	.76
	(.02)	(.04)	(.06)	(.03)	(.04)	(.06)	(.08)
(4)	.92	.91	.92	.90	.83	.83	.56
	(.02)	(.02)	(.02)	(.03)	(.02)	(.02)	(.10)
				24 items			
(1)	.97	.97	.97	.96	.97	.93	.71
	(.00)	(.00)	(.00)	(.00)	(.00)	(.02)	(.29)
(2)	.96	.94	.94	.94	.94	.83	.81
	(.00)	(.01)	(.01)	(.01)	(.01)	(.10)	(.02)
(3)	.90	.82	.84	.85	.81	.83	.64
	(.01)	(.03)	(.03)	(.03)	(.05)	(.07)	(.18)
(4)	.97	.96	.95	.95	.89	.92	.80
	(.01)	(.01)	(.03)	(.01)	(.03)	(.01)	(.02)

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed. The number in the parenthesis is the standard deviation of five replicated correlations.

Table 6
Means and S.D.s of RMSDs between True and Recovered Person Parameters for Case II (N=100)

Dist.	IRT-GRM	FA-PR	FA-PL	SSI	WMDU	IMDU	EMDU
				12 items			
(1)	.31	.32	.30	.33	.33	.55	.90
	(.03)	(.04)	(.02)	(.02)	(.02)	(.11)	(.30)
(2)	.33	.39	.39	.42	.39	.52	.67
	(.02)	(.03)	(.03)	(.04)	(.03)	(.06)	(.15)
(3)	.46	.61	.64	.59	.60	.69	.68
	(.04)	(.07)	(.09)	(.06)	(.06)	(.09)	(.11)
(4)	.39	.42	.40	.43	.58	.58	.93
	(.06)	(.06)	(.06)	(.06)	(.04)	(.04)	(.11)
				24 items			
(1)	.24	.23	.25	.27	.25	.38	.71
	(.01)	(.01)	(.01)	(.01)	(.02)	(.04)	(.32)
(2)	.26	.35	.36	.35	.35	.57	.62
	(.01)	(.02)	(.02)	(.02)	(.03)	(.15)	(.04)
(3)	.44	.59	.56	.55	.62	.56	.83
	(.03)	(.05)	(.05)	(.05)	(.07)	(.13)	(.20)
(4)	.26	.28	.31	.30	.46	.40	.63
	(.03)	(.03)	(.08)	(.03)	(.07)	(.03)	(.04)

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed. The number in the parenthesis is the standard deviation of five replicated RMSDs.

Table 7
Means and S.D.s of Correlations between True and Recovered Person Parameters for Case III (N=30)

#Items	Dist.	SSI	WMDU	IMDU	EMDU
12	(1)	.93	.92	.63	.56
		(.02)	(.05)	(.24)	(.22)
	(2)	.89	.88	.68	.64
		(.05)	(.05)	(.25)	(.11)
	(3)	.79	.76	.50	.71
		(.08)	(.04)	(.15)	(.07)
	(4)	.91	.84	.67	.55
		(.03)	(.07)	(.37)	(.07)
24	(1)	.96	.96	.89	.74
		(.01)	(.02)	(.06)	(.09)
	(2)	.94	.92	.56	.75
		(.01)	(.02)	(.19)	(.09)
	(3)	.87	.81	.44	.69
		(.04)	(.04)	(.26)	(.05)
	(4)	.94	.91	.88	.79
		(.02)	(.02)	(.09)	(.05)

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed. The number in the parenthesis is the standard deviation of five replicated correlations.

BEST COPY AVAILABLE

Table 8
Means and S.D.s of RMSDs between True and Recovered Person Parameters for Case III (N=30)

#Items	Dists.	SSI	WMDU	IMDU	EMDU
12	(1)	.13	.39	.80	.90
		(.03)	(.10)	(.26)	(.22)
	(2)	.21	.47	.75	.82
		(.10)	(.09)	(.28)	(.12)
	(3)	.40	.68	.97	.74
		(.16)	(.06)	(.14)	(.09)
	(4)	.18	.54	.72	.93
		(.06)	(.12)	(.38)	(.07)
24	(1)	.07	.28	.46	.70
		(.03)	(.08)	(.13)	(.12)
	(2)	.12	.38	.91	.69
		(.03)	(.06)	(.19)	(.14)
	(3)	.26	.60	1.03	.77
		(.07)	(.06)	(.23)	(.06)
	(4)	.11	.41	.46	.63
		(.04)	(.06)	(.16)	(.07)

Note: Dist.(1) = Normally distributed; Dist.(2) = Moderately skewed; Dist.(3) = Highly skewed; Dist.(4) = Differentially skewed. The number in the parenthesis is the standard deviation of five replicated RMSDs.